

(Translation)

PATENT OFFICE
JAPANESE GOVERNMENT

This is to certify that the annexed is a true copy of the
following application as filed with this Office

Date of Application.	September 30, 1999
Application Number	Japanese Patent Application No 277918 1999
Applicant(s)	Hitachi Software Engineering Co., Ltd

August 18, 2000

Commissioner.
Patent Office

Kozo Oikawa (seal)

Certificate No 2000-3066086



日 本 国 特 許 庁

PATENT OFFICE
JAPANESE GOVERNMENT

JCS11 U.S. PRO

09/677042



別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日

Date of Application:

1 9 9 9 年 9 月 3 0 日

出 願 番 号

Application Number:

平成 1 1 年特許願第 2 7 7 9 1 8 号

出 願 人

Applicant (s):

日立ソフトウェアエンジニアリング株式会社

2 0 0 0 年 8 月 1 8 日

特許庁長官
Commissioner,
Patent Office

及 川 耕 造

出証番号 出証特 2 0 0 0 - 3 0 6 6 0 8 6

【書類名】 特許願

【整理番号】 SK11A031

【提出日】 平成11年 9月30日

【あて先】 特許庁長官 殿

【国際特許分類】 G06F 19/00

【発明の名称】 遺伝子発現パターン表示方法および装置

【請求項の数】 5

【発明者】

【住所又は居所】 神奈川県横浜市中区尾上町 6 丁目 8 1 番地 日立ソフト
ウェアエンジニアリング株式会社内

【氏名】 野崎 康行

【発明者】

【住所又は居所】 神奈川県横浜市中区尾上町 6 丁目 8 1 番地 日立ソフト
ウェアエンジニアリング株式会社内

【氏名】 中重 亮

【発明者】

【住所又は居所】 神奈川県横浜市中区尾上町 6 丁目 8 1 番地 日立ソフト
ウェアエンジニアリング株式会社内

【氏名】 渡辺 恒彦

【特許出願人】

【識別番号】 000233055

【氏名又は名称】 日立ソフトウェアエンジニアリング株式会社

【代理人】

【識別番号】 100083552

【弁理士】

【氏名又は名称】 秋田 収喜

【電話番号】 03-3893-6221

【手数料の表示】

【予納台帳番号】 014579

【納付金額】 21,000円

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【ブルーフの要否】 要

【書類名】 明細書

【発明の名称】 遺伝子発現パターン表示方法および装置

【特許請求の範囲】

【請求項 1】 時間経過に伴って発現の度合いが変化する複数の遺伝子の時系列発現パターンを視覚的に表示する遺伝子発現パターン表示方法であって、

前記複数の遺伝子の時系列発現パターンデータの任意の時間区間を指定するステップと、

指定された時間区間における時系列発現パターンデータを予め定めた基準値によってクラスタリングし、さらに同一クラスタ内において基準値を変えながら正または負の時間方向にクラスタリングを繰り返し行ない、その結果を予め定めた表示形式で表示するステップと

を備えることを特徴とする遺伝子発現パターン表示方法。

【請求項 2】 前記基準値は、異なる遺伝子において発現のパターンが同じまたは異なるとみなすべき値であることを特徴とする請求項 1 記載の遺伝子発現パターン表示方法。

【請求項 3】 前記時間区間において、異なる 2 つ以上の遺伝子が、初め同じ発現パターンを示し、途中から異なる発現パターンを示すものを予め定めた表示形式で表示することを特徴とする請求項 1 または 2 記載の遺伝子発現パターン表示方法。

【請求項 4】 前記時間区間において、異なる 2 つ以上の遺伝子が、初め異なる発現パターンを示し、途中から同じ発現パターンを示すものを予め定めた表示形式で表示することを特徴とする請求項 1 または 2 記載の遺伝子発現パターン表示方法。

【請求項 5】 時間経過に伴って発現の度合いが変化する複数の遺伝子の時系列発現パターンデータをデータベースから取得し、時系列発現パターンを表示装置画面に視覚的に表示する遺伝子発現パターン解析装置であって、

前記データベースから取得した前記複数の遺伝子の時系列発現パターンデータの任意の時間区間を指定する入力手段と、

指定された時間区間における時系列発現パターンデータを予め定めた基準値に

よってクラスタリングし、さらに同一クラスタ内において基準値を変えながら正または負の時間方向にクラスタリングを繰り返し行ない、その結果を予め定めた表示形式で前記表示装置画面に表示させる演算手段とを備えることを特徴とする遺伝子発現パターン解析装置。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明は、特定の遺伝子とハイブリダイズさせることによって得られた時系列の遺伝子発現パターンを視覚的に分かり易く、そして遺伝子の機能・役割が推測し易い表示形式（または出力形式）によって表示するための表示方法および装置に関するものである。

【0002】

【従来の技術】

ゲノム配列が決定された種の増加に伴い、進化に対応するとみられる遺伝子を見つけ出し、どの生物にも共通に持っていると考えられる遺伝子の集合を探したり、それから逆に種に個別な特徴を推測するなど、種間の遺伝子の違いから何かを見出そうとする、いわゆるゲノム比較法が盛んに行われてきた。

【0003】

しかし近年、DNAチップやDNAマイクロアレイなどのインフラストラクチャの発達によって、分子生物学の興味は、種間の情報から種内の情報へ、すなわち同時発生解析へと移りつつあり、これまでの種間の比較と合わせて、情報の抽出から関連付けの場が大きく広がりを持ち始めている。

【0004】

例えば、既知の遺伝子と同一の発現パターンを示す未知の遺伝子が見つければ、それが既知の遺伝子と同様の機能があると類推できる。これら遺伝子や蛋白質そのものの機能的な意味付けは、機能ユニットや機能グループといった形で研究されている。またそれらの間の相互作用も、既知の酵素反応データや物質代謝データとの対応付けによって、あるいはより直接的に、ある遺伝子を破壊あるいは過剰反応させ、その遺伝子の発現をなくすか、あるいは多量に発現させ、その遺

伝子の直接的および間接的影響を、全遺伝子の発現パターンを調べることによって解析している。

【0005】

この分野において成功した事例として、スタンフォード大学のP.Brownらのグループによるイースト菌の発現解析が挙げられる (Michel B.Eisen et. al.: Cluster analysis and display of genome-wide expression patterns: Proc.Natl .Acad.Sci.(1998) Dec 8;95(25):14863-8)。彼らは、DNAマイクロアレイを用いて、細胞から抽出した遺伝子を時系列にハイブリダイズさせ、遺伝子の発現の度合い (ハイブリダイズした蛍光シグナルの輝度) を数値化した。数値に色を対応させることで、遺伝子の個々の発現過程を分かり易く表示させている。このとき、細胞の一連のサイクルにおいて発現パターンの過程が近い遺伝子同士 (任意の時点での発現の度合いが近いもの同士) をクラスタリングしている。

【0006】

図13は、この手法に従って遺伝子の発現状態1300を表示した例を示す図であり、横方向に時間軸、縦方向に遺伝子を並べている。このような表示方法をとることで、共通のクラスタに属する遺伝子は、共通の機能的性質をもつと類推することができる。なお、図13における1つ1つの枠1301が1つの遺伝子のある時刻における発現状態を示すものであり、図13では白黒の濃度を変えて発現状態を模式的に示している。

【0007】

【発明が解決しようとする課題】

ところが、実際の遺伝子間の発現過程では、細胞の全サイクルにおいて同様の発現パターンを持つ幾つかの遺伝子グループを見つけ出すことで、その細胞全ての遺伝子間の関連が解明されるというほど単純ではない。

【0008】

例えば、ある時点において異なる遺伝子が同じ機能のために同様に発現しているが、その後、次のある時点では別々の役割を持つような場合がある。当然この場合、遺伝子の発現過程は異なる。細胞の全サイクルにおいて発現のパターンが近いもの同士をクラスタリングして表示させる従来技術の手法では、これらの遺

伝子は別々のクラスタとして分類されるため、こういった性質を見つけ難いという難点があった。

【0009】

本発明は、このような従来技術の問題点を鑑み、ある時点において異なる遺伝子が同じ機能のために同様に発現しているが、ある時点では別々の役割を持つような場合を見つけ出し、これを効果的に表示することが可能な遺伝子発現パターン表示方法および装置を提供することを目的とする。

【0010】

【課題を解決するための手段】

本発明では、前記目的を達成するために、時間経過に伴って発現の度合いが変化する複数の遺伝子の時系列発現パターンを視覚的に表示する遺伝子発現パターン表示方法であって、

前記複数の遺伝子の時系列発現パターンデータの任意の時間区間を指定するステップと、指定された時間区間における時系列発現パターンデータを予め定めた基準値によってクラスタリングし、さらに同一クラスタ内において基準値を変えながら正または負の時間方向にクラスタリングを繰り返し行い、その結果を予め定めた表示形式で表示するステップとを備えることを特徴とする。

【0011】

また、前記時間区間において、異なる2つ以上の遺伝子が、初め同じ発現パターンを示し、途中から異なる発現パターンを示すものを予め定めた表示形式で表示することを特徴とする。

【0012】

また、前記時間区間において、異なる2つ以上の遺伝子が、初め異なる発現パターンを示し、途中から同じ発現パターンを示すものを予め定めた表示形式で表示することを特徴とする。

【0013】

また、時間経過に伴って発現の度合いが変化する複数の遺伝子の時系列発現パターンデータをデータベースから取得し、時系列発現パターンを表示装置画面に視覚的に表示する遺伝子発現パターン解析装置であって、

前記データベースから取得した前記複数の遺伝子の時系列発現パターンデータの任意の時間区間を指定する入力手段と、指定された時間区間における時系列発現パターンデータを予め定めた基準値によってクラスタリングし、さらに同一クラスタ内において基準値を変えながら正または負の時間方向にクラスタリングを繰り返し行い、その結果を予め定めた表示形式で前記表示装置画面に表示させる演算手段とを備えることを特徴とする。

【0014】

【発明の実施の形態】

以下、図面を参照して本発明の実施の形態を説明する。

図1は、本発明の遺伝子発現パターン表示方法を適用した遺伝子発現パターン解析装置の一実施形態を示すシステム構成図である。この実施形態の解析装置は、一連の細胞のプロセスにおいて遺伝子の発現の度合いを数値化した遺伝子発現パターンデータを格納した記憶装置（またはデータベース）101、発現パターンデータを視覚化して表示するための表示装置102、本システムへの値の入力や選択の操作を行なうためのキーボード103およびマウス104、遺伝子の発現過程に応じて発現パターンデータのクラスタリングを行なうクラスタリング処理部105から構成される。このクラスタリング処理部105は、コンピュータとそのプログラムによって具体化されるものである。

【0015】

ここで、記憶装置101に代えて、ネットワーク等を介して遠隔地に設置されたサーバコンピュータが管理しているデータベースから遺伝子発現パターンデータを取得する構成にする実施形態がある。

【0016】

本実施形態においては、細胞の一連のサイクルにおいて特定の時間区間を指定し、その時間区間において細かい粒度でクラスタリングを行なう。

【0017】

すなわち、同一のクラスタに属する遺伝子は1つに束ね、異なるクラスタとの間には線を引き、さらに、クラスタ内の遺伝子において更にクラスタリングを行なう。細かい粒度のクラスタリングを範囲の始めから正の時間方向へ繰り返し行

なうと、図2に示すように、遺伝子の発現過程が木構造のように分岐して表現できる。図2において、201は、指定された時間区間、すなわちクラスタリング範囲である。

【0018】

これは、指定された時間区間の始めにおいて同じ発現パターンを示し、時間区間の途中で異なる発現パターンを示したことを意味している。このような表示が得られた場合、始めの時点では異なる遺伝子が同じ機能のために同様に発現しているが、ある時点において別々の役割を持つため異なって発現したと類推することができる。

【0019】

同様に、細かい粒度のクラスタリングを範囲の終端から負の時間方向へ繰り返し行なうと、遺伝子の発現の過程が、図3のように、逆の木構造のような分岐構造として表現することができる。

【0020】

これは、範囲の始めにおいて異なる発現パターンを示し、範囲の途中で同じ発現パターンを示したことを意味している。このような表示が得られた場合、始めの時点では異なる遺伝子が異なる機能を持っていたが、ある時点において同様の役割を持ったと類推することができる。

【0021】

図4は、遺伝子の発現パターンデータをクラスタリングして表示するクラスタリング処理部105におけるアルゴリズムの概要を示すフローチャートである。

【0022】

ここではまず、初期パラメータを設定し（ステップ401）、表示位置決定処理を行なう（ステップ402）。初期パラメータについては、後述する。その後、表示処理を行ない、処理を終了する（ステップ403）。本アルゴリズムは、図2に示したように、異なる遺伝子が、ある時間区間において、始めにおいて同じ発現パターンを示し、途中で異なる発現パターンを示したことを表示するものである。

【 0 0 2 3 】

図 5 は、本アルゴリズムで使われる変数と実データとの対応関係を示す説明図である。図 6 は、図 4 中の初期パラメータ設定処理（ステップ 4 0 1）に関するアルゴリズムの詳細を示している。

【 0 0 2 4 】

まず、遺伝子発現パターンデータを記憶装置 1 0 1 から読み込む。この遺伝子発現パターンデータには、図 5 に示すように $m + 1$ 個のサンプル遺伝子 g_0, g_1, \dots, g_m について、時刻 T_0, T_1, \dots, T_n において実験した結果の発現パターンデータが入っているものとする。そこで、時刻 T_j における遺伝子 g_i の発現の観測値を $g[j][i]$ とおく（ステップ 6 0 1）。

【 0 0 2 5 】

次に、キーボード 1 0 3、マウス 1 0 4 を使って、クラスタリング適用範囲（開始時刻 T_{start} 、終了時刻 T_{end} ）、異なるクラスタとみなすべき基準を示す正数値（ $K_{\text{start}}, K_{\text{start}+1}, \dots, K_{\text{end}}$ ）、クラスタリングの粒度を示す整数（ S ）、クラスタリング手法をそれぞれ入力する（ステップ 6 0 2）。

【 0 0 2 6 】

クラスタリング適用範囲とは、図 2、図 3 に太枠実線 2 0 1 で示すように、細胞の一連のプロセスにおいて、より詳しくクラスタリングする時間区間を示す。例えば細胞の一連のプロセスにおいて、ある時刻で細胞に特殊な発現パターンがみられた場合、その時刻の前後をクラスタリング適用範囲に指定することで、全遺伝子の発現状態をより詳しくモニタリングするように選択する。従来のクラスタリングとの基本的な相違点は、図 1 3 のような細胞の全プロセスにおいて発現状態の近いもの同士をクラスタリングするのではなく、図 2 に示すような相異なる遺伝子が範囲の始めにおいて同じ発現パターンを示し、範囲の途中で異なる発現パターンを示したことを表示するところにある。

【 0 0 2 7 】

異なるクラスタとみなす基準とは、異なるクラスタの間の非類似度が最低でもどれくらいの値をとるかを示すものである。すなわち、クラスタ間の閾値 K を示している。閾値が $K_{\text{start}}, K_{\text{start}+1}, \dots, K_{\text{end}}$ と可変に設定できることで、時間

によって粗いクラスタリングから細かいクラスタリングまで調節できる。

【0028】

また、クラスタリングを行なうときの非類似度の計算において、本システムでは、発現データの時刻 T_0, T_1, \dots, T_n における全てのデータを非類似度の計算の対象とせずに、ある時間区間を設けて、その時間区間内におけるデータを非類似度の計算の対象とする。この時間区間を図5に示すようにスリット501、このスリット501の長さ（時間軸方向の幅） S をクラスタリングの粒度とよぶ。本アルゴリズムでは、まずスリット501の先頭を T_{start} に合わせてデータを T_{start} から $T_{start+S}$ の範囲でクラスタリングを行ない、そこで分割された各々のクラスタ内において、スリット501を時刻が正の方向へ1つずらし、 $T_{start+1}$ から $T_{start+S+1}$ の範囲でクラスタリングを行なう。このような操作をスリットの後端が T_{end} になるまで逐次実行する。したがって、粒度が細かいほど、すなわち時間区間の幅が短いほど、より細かい遺伝子間の発現の違いを表すことができる。

【0029】

クラスタリング手法では、クラスタリングにおいて個体同士の相関関係を表す類似度または非類似度（ピアソンの相関係数、ユークリッド平方距離、標準化ユークリッド平方距離、マハラノビスの距離、ミンコフスキー距離など）及びクラスタ合併のアルゴリズム（最短距離法、最長距離法、群平均法、重心法、メディアン法、ワード法、可変法など）を指定する。本アルゴリズムは非類似度を対象としているが、クラスタリング手法において類似度を選択した場合は、計算した類似度に負符号を付けたり、逆数をとるなどの操作を施し、非類似度に変換すればよい。

【0030】

これらの値を設定したら、それぞれの項目が正しいかどうか調べる。クラスタリング適用範囲 T_{start} 、 T_{end} が T_0 から T_n の範囲に含まれているか（ステップ603）、クラスタリングの粒度 S がクラスタリング適用範囲の幅を超えてないか（ $S \leq end - start$ ）（ステップ604）、また設定したクラスタリング手法において、合併アルゴリズムを重心法、メディアン法、ワード法を選択した時

、非類似度においてユークリッド平方距離を選択したかなど、類似度または非類似度と合併アルゴリズムは妥当な組み合わせか（ステップ 6 0 6）を調べる。もし、これらの値で正当なものが入っていないならば、表示装置 1 0 2 にエラーを出力し、再入力を促す（ステップ 6 0 7）。

【 0 0 3 1 】

しかし、設定項目が適切であった場合、次に、 $i=1,2,\dots,m$ に対して遺伝子 g_i の平均発現度 $G_i = (g[0][i] + g[1][i] + \dots + g[n][i]) / n$ を求める（ステップ 6 0 8）。

【 0 0 3 2 】

次に、個々の遺伝子の表示情報を格納するために図 5 に示すような配列 $l[I]$ ($I=0,1,\dots,m$) 5 0 2 と整数値変数 $lmax$ を用意する。各 $l[I]$ は構造体データで、図 5 に示すように遺伝子のインデックスを表すメンバ (index) と異なるクラスタ間の仕切り線の位置を表すメンバ (linepos) からなる。構造体のメンバは、 $l[I].index, l[I].linepos$ という形で値を代入・参照できる。そこで、全ての I に対して $l[I].linepos$ の値を T_{end} として初期化し（ステップ 6 0 9）、さらに $lmax$ の値を「0」としておく（ステップ 6 1 0）。次に、変数 t に start の値を設定する（ステップ 6 1 1）。

【 0 0 3 3 】

本アルゴリズムでは、整数値の集合を表す“クラスタ”と呼ばれる抽象データ型を使っている。クラスタには、整数の登録、削除、登録データの参照のインタフェースを備えているものとする。

【 0 0 3 4 】

クラスタ B を生成し、そこに $\{0,1,2,\dots,m\}$ を登録し処理を終了する（ステップ 6 1 2）。

【 0 0 3 5 】

以上のように初期設定をした後、クラスタリング適用範囲 2 0 1 に対して処理を行なう。すなわち、上で定めた t と B とを引数として表示位置決定処理（図 4 のステップ 4 0 2 の処理 A）を行なう。

【0036】

図7は、図4中の表示位置決定処理（処理A）の詳細を示すフローチャートであり、この処理Aの中で配列lに表示情報を登録する。

【0037】

まず、引数として渡されたクラスタをB、時刻をtとする（ステップ701）。ここでBを更にクラスタリングする（処理B）。このときtとBを引数として与える。処理Bの結果として、総クラスタ数がcmaxに、クラスタリング結果がA[J]（ $J=1, 2, \dots, cmax$ ）に設定される（ステップ702）。処理Bの詳細については後述する。

【0038】

次に、「 $t + S$ 」がendと等しいかどうか判定する（ステップ703）。endの時はスリット501の終端がクラスタリング適用範囲201の終わりに来たことを意味し、ここでクラスタリング処理を終了する。このとき、 $J=1$ としてJがcmaxを超えるまで、各々のクラスタに対して次の処理を実行する（ステップ704, 705）。クラスタA[J]の要素が $\{i_1, \dots, i_k\}$ であるとき、これらの要素を一定の基準の下に並べて表示する。ここでは各要素に対応する遺伝子の平均発現度 G_{i1}, \dots, G_{ik} を値の降順に並べて、それを G_{j1}, \dots, G_{jk} とおく（ステップ706）。

【0039】

次に配列lの値を入力する。すなわち、発現パターンデータの位置情報を表すl[].indexに平均輝度が降順になるように $l[lmax].index=j_1$ 、 $l[lmax+1].index=j_2$ 、 \dots 、 $l[lmax+k-1].index=j_k$ と設定し（ステップ707）、異なるクラスタとの仕切り線（図2の202で代表して示す横方向の太実線）を表す $l[lmax+k-1].linepos$ に時刻tから $t+S (= T_{end})$ の範囲まで線を引くことを示すtの値を入力する（ステップ708）。

【0040】

次に、配列lの入力済みデータの最大数を示すlmaxにkを加算する（ステップ709）。次に、Jを1つインクリメントし、次のクラスタの処理に移る（ステップ710）。

【0041】

一方、ステップ703において、「 $t+S$ 」が end と一致しない場合、すなわちスリット501の終端がクラスタリング適用範囲201の終わりに来ていないとき、 t を1つインクリメントし、 J に「1」を設定する（ステップ711）。 J が c_{max} を超えるまで、各々のクラスタに対して次の処理を行なう（ステップ712）。すなわち B に $A[J]$ を代入し（ステップ713）、引数として時刻 t 、クラスタ B を与えて表示位置決定処理（処理A）を行なう（ステップ714）。次に、異なるクラスタとの仕切り線を表す $l[l_{max}-1].linepos$ に時刻 t から T_{end} の範囲まで線を引くことを示す t を入力する（ステップ715）。そして、 J を1つインクリメントし、次のクラスタの処理に移る（ステップ716）。全てのクラスタ $A[J]$ ($J=1, \dots, c_{max}$) に関する処理が終われば終了する。

【0042】

図8および図9は、クラスタリング処理（処理B）のアルゴリズムを示すフローチャートである。

まず、引数として入力されたクラスタを B 、入力された時刻を t に入れる（ステップ801）。

【0043】

次に、クラスタ B の要素が i_1, \dots, i_k であるとき、 i_1, \dots, i_k に対応する遺伝子間の時刻 t から時刻 $t+S$ における類似度または非類似度 d_{ij} ($i < j$ かつ $i, j \in \{i_1, i_2, \dots, i_k\}$)を求める（ステップ802）。

【0044】

ここで、遺伝子 g_i, g_j に対する遺伝子発現データ $\{g[0][i], g[1][i], \dots, g[n][i]\}$ 、 $\{g[0][j], g[1][j], \dots, g[n][j]\}$ の時刻 t から時刻 $t+S$ における類似度（非類似度）とは、例えば以下のような計算で求める量である（ステップ802）。

【0045】

(1) 類似度としてピアソンの相関係数を指定したとき

【0046】

【数1】

$$d_{i,j} = \frac{\sum_{k=1}^{t+s} (g[k][i] - \overline{g[i]})(g[k][j] - \overline{g[j]})}{\sqrt{\left\{ \sum_{k=1}^{t+s} (g[k][i] - \overline{g[i]})^2 \right\} \left\{ \sum_{k=1}^{t+s} (g[k][j] - \overline{g[j]})^2 \right\}}} \quad \dots\dots (1)$$

$$\text{ただし } \overline{g[l]} = \frac{1}{s} \sum_{k=1}^{t+s} g[k][l]$$

【0047】

となる。本アルゴリズムでは非類似度を対象にしているので、類似度を適用する場合には負符号を付ける、逆数をとるなどの操作をして非類似度に変換しなければならない。

【0048】

(2) 非類似度としてユークリッド平方距離を指定したとき、

【0049】

【数2】

$$d_{i,j} = \sum_{k=1}^{t+s} (g[k][i] - g[k][j])^2 \quad \dots\dots (2)$$

【0050】

(3) 標準化ユークリッド平方距離を指定したとき、

【0051】

【数3】

$$d_{i,j} = \sum_{k=1}^{t+s} (g[k][i] - g[k][j])^2 / s_k^2 \quad \dots\dots (3)$$

ただし s_k^2 は変数 $g[k][0], \dots, g[k][n]$ の分散。

【0052】

(4) マハラノビスの距離を指定したとき、

【0053】

【数4】

$$d_{i,j} = (g[i] - g[j])S^{-1}(g[i] - g[j]) \quad \dots\dots (4)$$

ただし $g[l] = (g[t][l], \dots, g[t+S][l])$ 、
 S は $g[i], g[j]$ の共分散行列。

【0054】

(5) ミンコフスキー距離を指定したとき、

【0055】

【数5】

$$d_{i,j} = \left\{ \sum_{l=1}^{t+S} |g[l][i] - g[l][j]|^k \right\}^{1/k} \quad \dots\dots (5)$$

【0056】

クラスタ $C[1], \dots, C[k]$ を生成し、それぞれのクラスタに $C[1] \leftarrow \{i_1\}, \dots, C[k] \leftarrow \{i_k\}$ を登録しておく（ステップ803）。そして、生成したクラスタの数を表す変数 $ccnt$ に k を代入しておく（ステップ804）。次に、空集合のクラスタ D を生成する（ステップ805）。

【0057】

次に、ここまでで計算した非類似度 $d_{i,j}$ ($i, j \in \{1, 2, \dots, ccnt\} - D$) の値の最小値 $d_{p,q}$ を求め、先に設定した閾値 K_t 以下かどうか判定する（ステップ806、807）。 $d_{p,q}$ が K_t 以下のとき次のことを実行する。クラスタ $C[ccnt+1]$ を新たに生成し、クラスタ $C[p]$ とクラスタ $C[q]$ に含まれる要素の和集合をクラスタ $C[ccnt+1]$ に登録し（ステップ808）、クラスタ $C[p]$ とクラスタ $C[q]$ に含まれる要素を削除する（ステップ809）。次に、 $C[p]$ と $C[q]$ はもう必要ないので、 D に p, q を登録する（ステップ810）。次に、クラスタ $C[h]$ ($h \in \{1, 2, \dots, ccnt\} - D$) とクラスタ $C[ccnt+1]$ 間の時刻 t から時刻「 $t+S$ 」における非類似度 $d_{h,ccnt+1}$ を求める（ステップ811）。ここで $d_{h,ccnt+1}$ は、次の計算式で求めることができる。すなわち

【0058】

【数6】

$$d_{h,ccnt+1} = \alpha d_{h,p} + \beta d_{h,q} + \gamma d_{p,q} + \delta |d_{h,p} - d_{h,q}| \quad \dots\dots (6)$$

【0059】

ここで α 、 β 、 γ 、 δ は、 $n(k)$ をクラスタC[k]内の要素の個数としたとき、クラスタリング手法が

- (1) 最短距離法するとき $\alpha=0.5$ 、 $\beta=0.5$ 、 $\gamma=0$ 、 $\delta=-0.5$
 - (2) 最長距離法するとき $\alpha=0.5$ 、 $\beta=0.5$ 、 $\gamma=0$ 、 $\delta=0.5$
 - (3) 群平均法するとき $\alpha=n(p)/n(ccnt+1)$ 、 $\beta=n(q)/n(ccnt+1)$ 、 $\gamma=0$ 、 $\delta=0$
 - (4) 重心法するとき $\alpha=n(p)/n(ccnt+1)$ 、 $\beta=n(q)/n(ccnt+1)$ 、
 $\gamma=-n(p)n(q)/n(ccnt+1)^2$ 、 $\delta=0$
 - (5) メディアン法するとき $\alpha=0.5$ 、 $\beta=0.5$ 、 $\gamma=-0.25$ 、 $\delta=0$
 - (6) ウォード法するとき $\alpha=\{n(h)+n(p)\}/\{n(h)+n(ccnt+1)\}$ 、
 $\beta=\{n(h)+n(q)\}/\{n(h)+n(ccnt+1)\}$ 、 $\gamma=-n(h)/\{n(h)+n(ccnt+1)\}$ 、 $\delta=0$
- である。

【0060】

次に、生成したクラスタの数を表す変数ccntに「1」を加える（ステップ812）。これらの処理を更新した $d_{i,j}$ ($i, j \in \{1, 2, \dots, ccnt\} - D$)の最小値が K_t より大きくなるまで続ける。

【0061】

ステップ807において $d_{i,j}$ の最小値 $d_{p,q}$ が K_t より大きいとき、クラスタリングを終えて、結果の出力処理を行なう。まず、クラスタC[1]からC[ccnt]で、空集合でないものを判定し、この総数をcmaxに入力する（ステップ813）。そして、cmax個のクラスタA[1], ..., A[cmax]を生成する（ステップ814）。空集合でないクラスタに対し、それに含まれる遺伝子の平均発現度の平均をとる。すなわち、クラスタC[p] = { i_1, \dots, i_k }に対して、 $G'_p = (G_{i_1} + \dots + G_{i_k}) / k$ を求める。この値を降順に並べたものを、 $G'_{p1}, \dots, G'_{pcmax}$ としたときA[1] ← C[p₁], ..., A[cmax] ← C[p_{cmax}]を登録する（ステップ815）。最後に

、総クラスタ数 c_{\max} とクラスタ $A[1], \dots, A[c_{\max}]$ を出力し（ステップ 8 1 6）
、処理を終了する。

【0 0 6 2】

図 1 0 は、図 4 における表示処理のアルゴリズムの詳細を示すフローチャートである。このアルゴリズムは、配列 $l[i]$ を読み込み、対応する遺伝子の発現データを表示する処理である。

【0 0 6 3】

まず i の値を「0」とし（ステップ 1 0 0 0）、 i の値が l_{\max} と等しくなるまで、各々の遺伝子発現データに対して以下の操作を続ける（ステップ 1 0 0 1）。次に、 $x=l[i].index$ が指す遺伝子 1 行分の発現データ $g[k][x]$ ($k=0, 1, \dots, n$)の数値を対応する表示色に置き換え、第 i 行として 1 行にわたり表示する（ステップ 1 0 0 2）。更に、クラスタ間の仕切り線を、今表示した第 i 行のすぐ下の時刻 $l[i].linepos$ から T_{end} の範囲に引く（ステップ 1 0 0 3）。

【0 0 6 4】

ここで、 $l[i].linepos$ の値が、初期値 T_{end} の場合は、クラスタ間の仕切り線は存在せず線も書く必要が無い。 i を 1 つずつインクリメントし（ステップ 1 0 0 4）、ステップ 1 0 0 1 において i が l_{\max} になったら、処理を終える。

【0 0 6 5】

以上の処理によって、図 2 に示したような、相異なる遺伝子がクラスタリング適用範囲の始めにおいて同じ遺伝子発現パターンを示し、範囲の途中で異なる発現パターンを示すような状況を効果的に表示することができる。

【0 0 6 6】

また、図 3 に示したような、相異なる遺伝子がクラスタリング適用範囲の始めにおいて異なる遺伝子発現パターンを示し、範囲の途中で同じ発現パターンを示すような状況を効果的に表示する場合には、ステップ 6 0 9（図 6）において $l[i].linepos$ に T_{start} を、ステップ 6 1 1 において t に end を設定し、ステップ 7 0 3（図 7）において $t+S=end$ の判定条件を $t-S=start$ にし、ステップ 7 1 1 において $t \leftarrow t+1$ を $t \leftarrow t-1$ に置き換え、ステップ 1 0 0 3（図 1 0）においてクラスタ間の仕切り線を、 T_{start} から $l[i].linepos$ の範囲に引けばよい。

これは、はじめスリットの終端部分を T_{end} に設定しておき、時間軸の負の方向へ 1 つずつスリットを移動してクラスタリングすることを意味している。

【0067】

また、これらの詳細なクラスタリング手法の応用例として、クラスタリング適用範囲の前方から時間軸の正の方向へスリットを動かしてクラスタリングを行ない、図 11 に示したような表示が得られた場合を考える。このとき、図 11 の点線 1101, 1102 で囲んだような似通った発現パターンが見られた場合、それらの遺伝子をマーキング (1103) しておき、クラスタリング適用範囲 201 の後方から時間軸の負の方向に向けてクラスタリングを行なう。もし、図 12 に示したようにマーキング (1103) した遺伝子が互いに近い位置にあるものが見つかる (例えば①と④、③と⑥など) ならば、これらの遺伝子は始め異なる遺伝子発現パターンを示し、途中で同じ発現パターンを示すことを意味しており、このような双方向のクラスタリングによって個々の遺伝子の発現状態を容易に推測することが出来る。

【0068】

更に、 T_{start} を T_0 に T_{end} を T_n に、スリット幅 S を n に設定すれば、従来の技術の中で説明した P. Brown らの結果と同様の表示を得ることが出来る。

【0069】

なお、本発明は、上記実施形態に限定されるものではなく、実施に際しては、細部を種々変更して実施することができる。例えば、途中から発現パターンが変わった部分あるいは境界においては、フリッカ表示、高輝度表示、色反転表示などの既知の表示形態を各種組み合わせる表示することができる。

【0070】

また、クラスタリング処理部 105 の処理は、プログラムとして CD-ROM 等の記録媒体に記録してコンピュータユーザに提供することができる。

【0071】

また、遺伝子のデータとしては、時系列の発現データに限定されるものではなく、図 3 または図 4 における横軸 (時間軸) を他の基準にとり変えることによって、例えば異なる実験間について比較を行うなどの利用が考えられる。

【0072】

また、解析結果を表示装置画面に表示する例を説明したが、最近においては多色プリンタの精度が向上しているため、多色プリンタで印刷出力する構成であってもよい。本発明の表示とは、プリンタで視覚的に印刷出力する概念を含むものである。

【0073】

【発明の効果】

以上説明したように、本発明によれば、細胞の発現サイクルの一部区間を指定し、その範囲において細かい粒度でクラスタリングを行なうことができる。そして、この表示結果に基づいて、利用者は遺伝子の発現経過の状態をより詳細に観測することができ、遺伝子の発現状態から生物学的機能を効率的よく推測することができる。

【図面の簡単な説明】

【図1】

本発明を適用した解析装置の一実施形態を示すシステム構成図である。

【図2】

クラスタリングの範囲を制限して細かい粒度でクラスタリングしたときの遺伝子発現パターン表示例（その1）を示す模式図である。

【図3】

クラスタリングの範囲を制限して細かい粒度でクラスタリングしたときの遺伝子発現パターン表示例（その2）を示す模式図である。

【図4】

クラスタリング処理の概要を示すフローチャートである。

【図5】

クラスタリング処理で使用する変数と実データの関係を示す説明図である。

【図6】

初期パラメータの設定に関するアルゴリズムを示すフローチャートである。

【図7】

表示位置決定処理のアルゴリズムを示すフローチャートである。

【図 8】

クラスタリングのアルゴリズムを示すフローチャートである。

【図 9】

図 8 の続きを示すフローチャートである。

【図 1 0】

表示処理のアルゴリズムの概要を示すフローチャートである。

【図 1 1】

クラスタリング適用範囲の前方から時間軸の正の方向へスリットを動かしてクラスタリングを行ったときの遺伝子発現パターン表示例を示す説明図である。

【図 1 2】

クラスタリング適用範囲の後方から時間軸の負の方向へスリットを動かしてクラスタリングを行ったときの遺伝子発現パターン表示例を示す説明図である。

【図 1 3】

細胞の全プロセスにおいて発現状態の近いものどうしをクラスタリングしたときの遺伝子発現パターン表示例を示す説明図である。

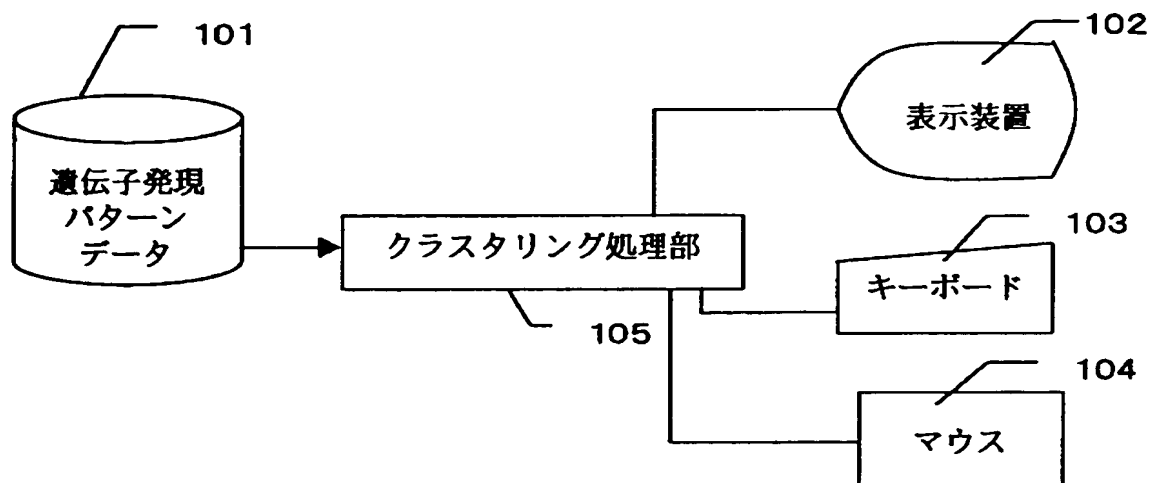
【符号の説明】

1 0 1 … 遺伝子発現パターンデータの記憶装置、1 0 2 … 表示装置、1 0 3 … キーボード、1 0 4 … マウス、1 0 5 … クラスタリング処理部、2 0 1 … クラスタリング範囲、5 0 1 … スリット。

【書類名】 図面

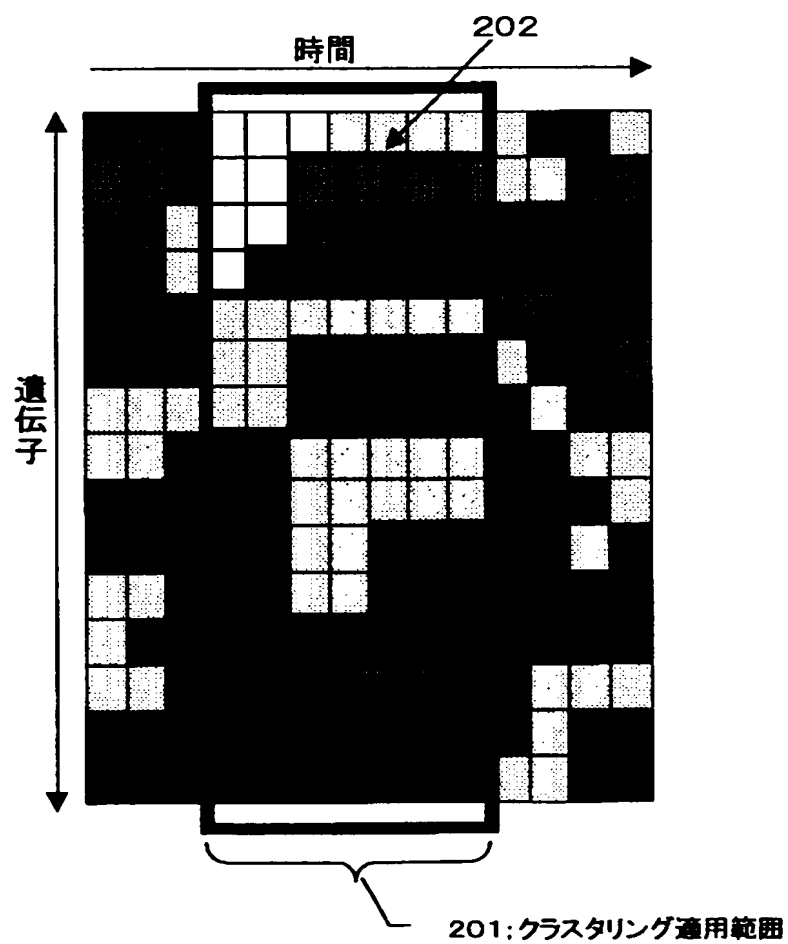
【図 1】

図 1



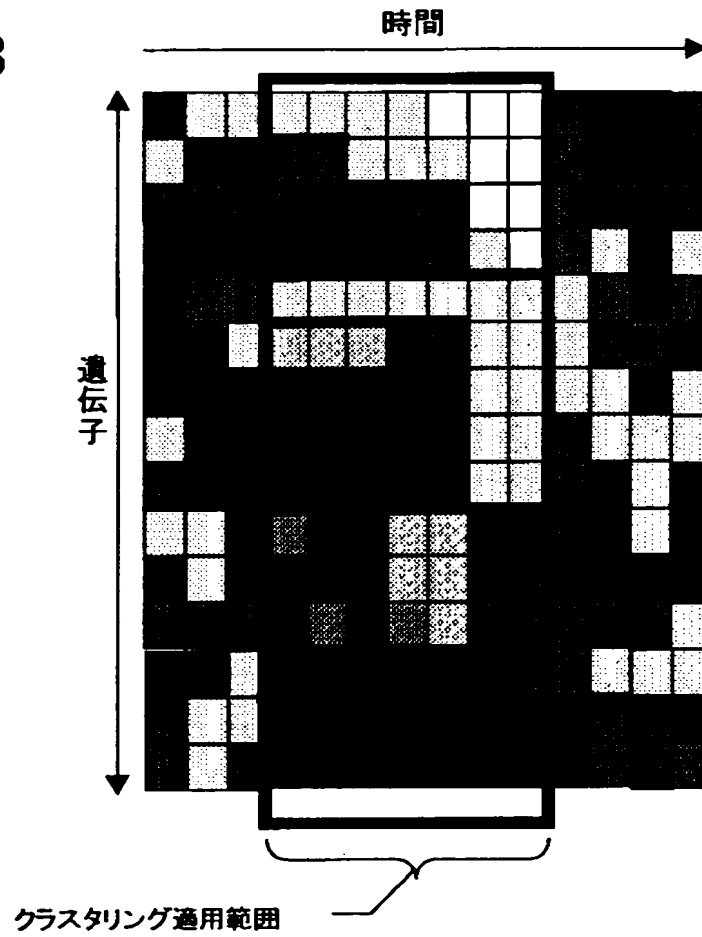
【図 2】

図 2



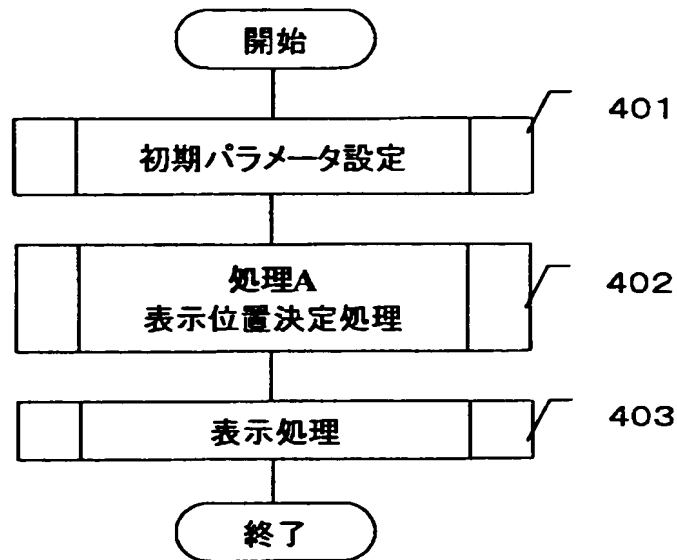
【図 3】

図 3



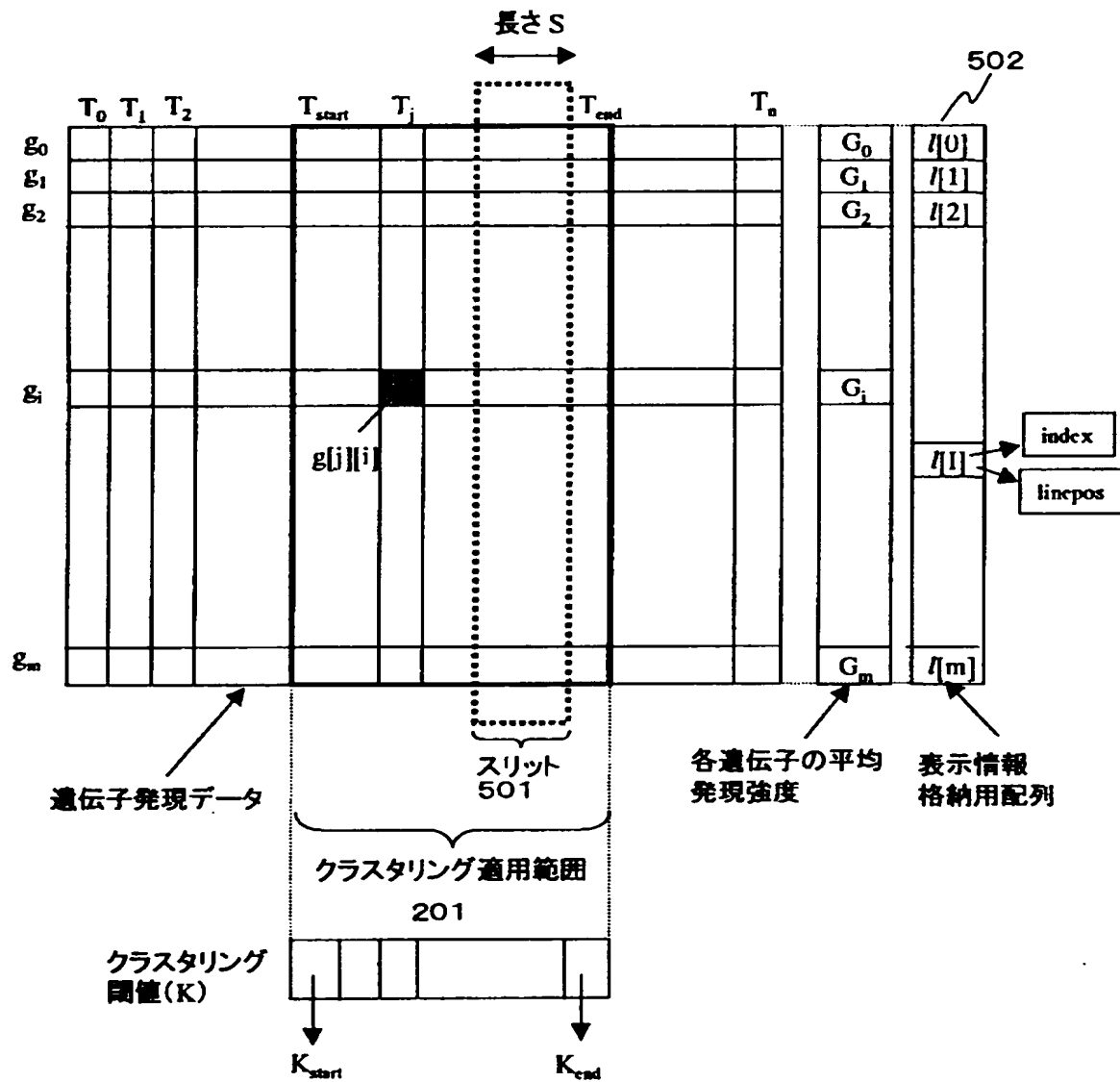
【図 4】

図 4

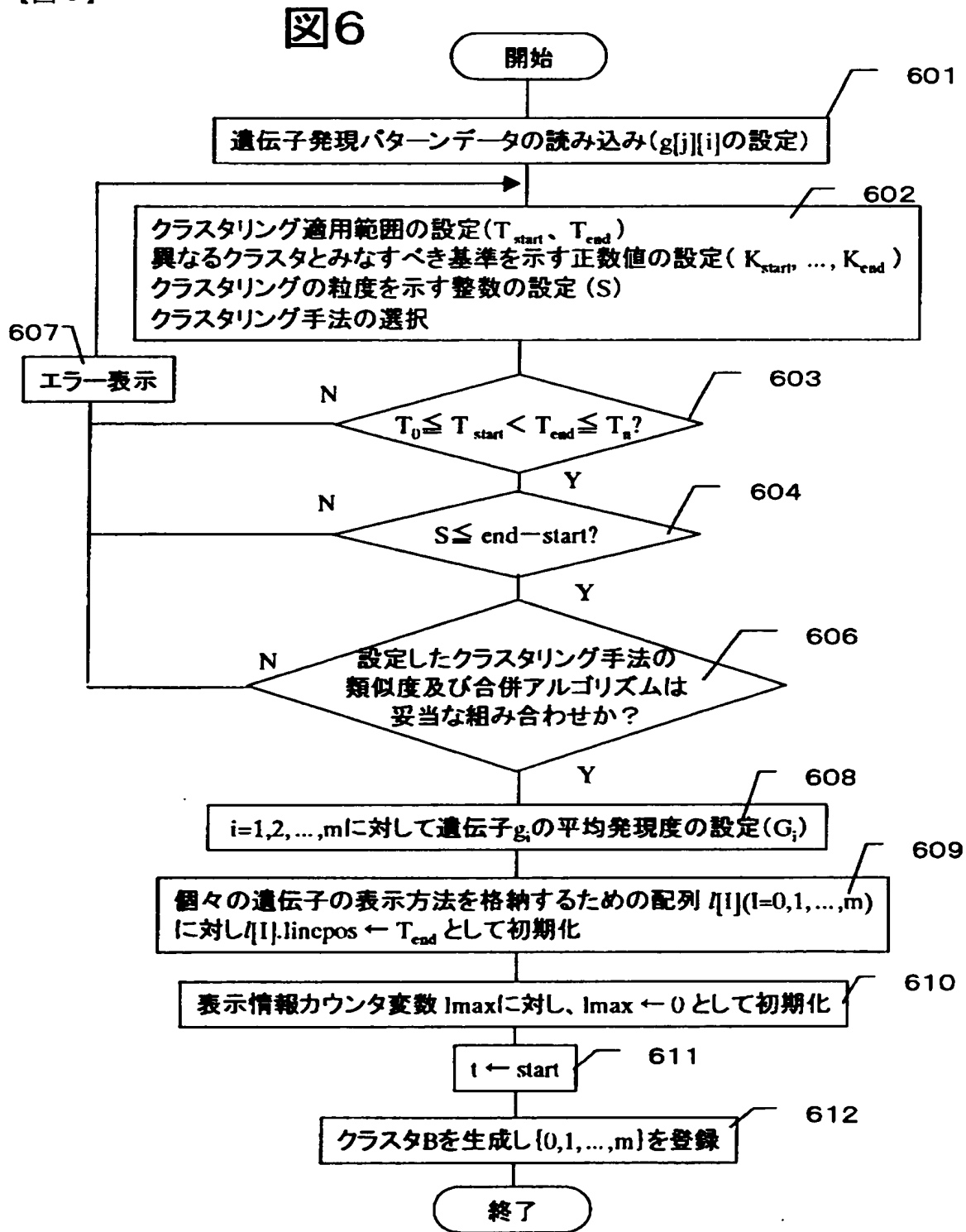


【図 5】

図 5

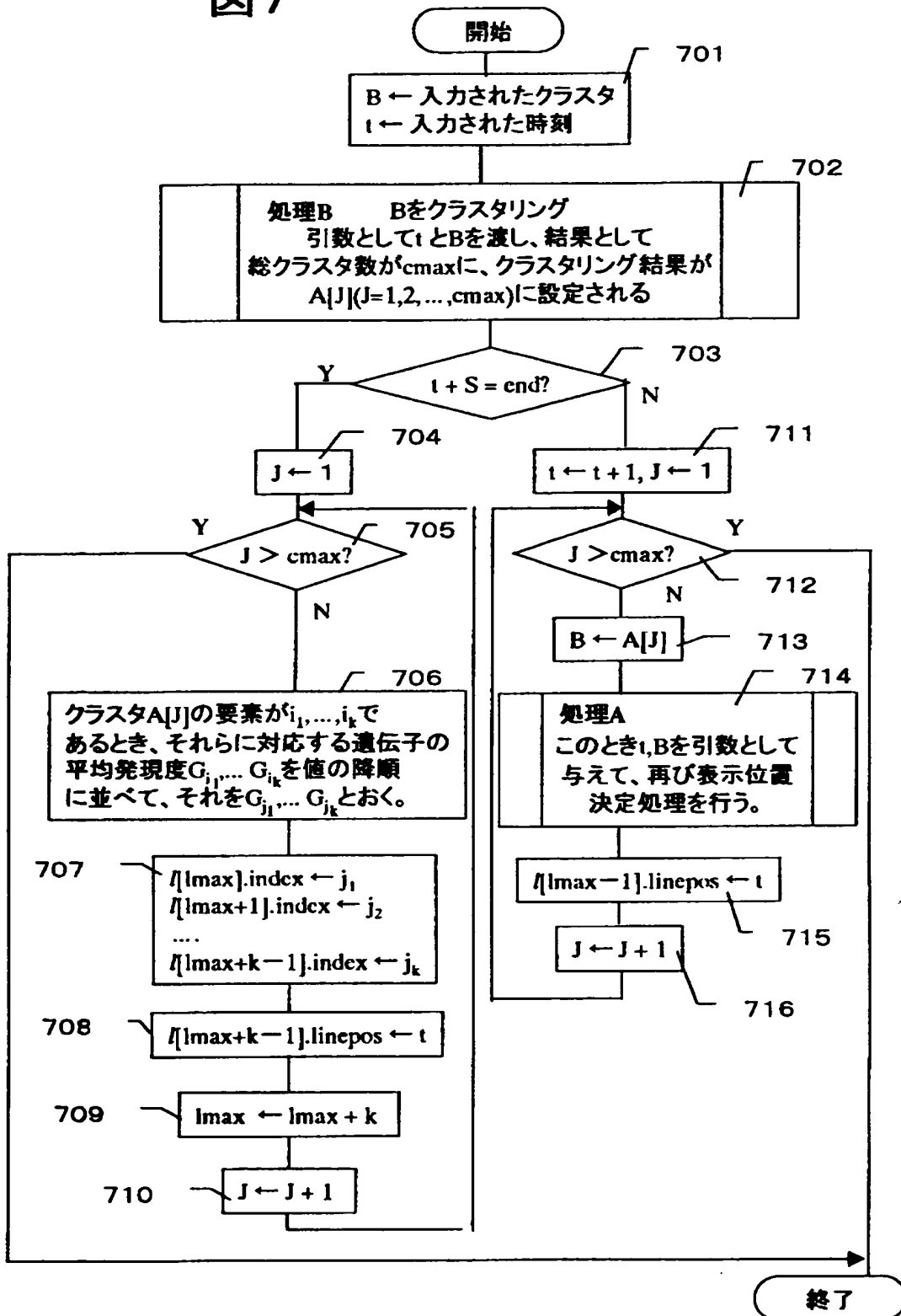


【図 6】



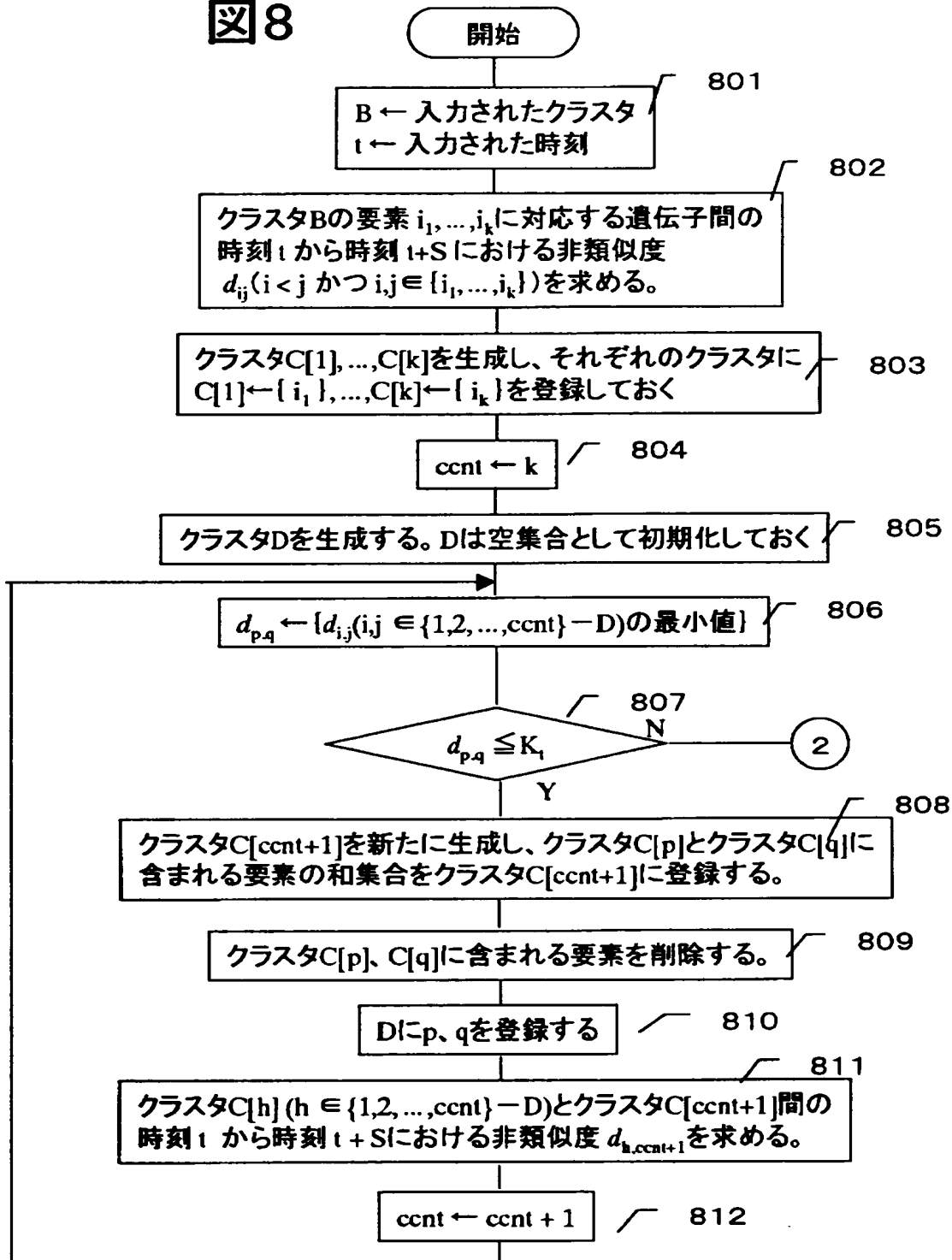
【図 7】

図 7



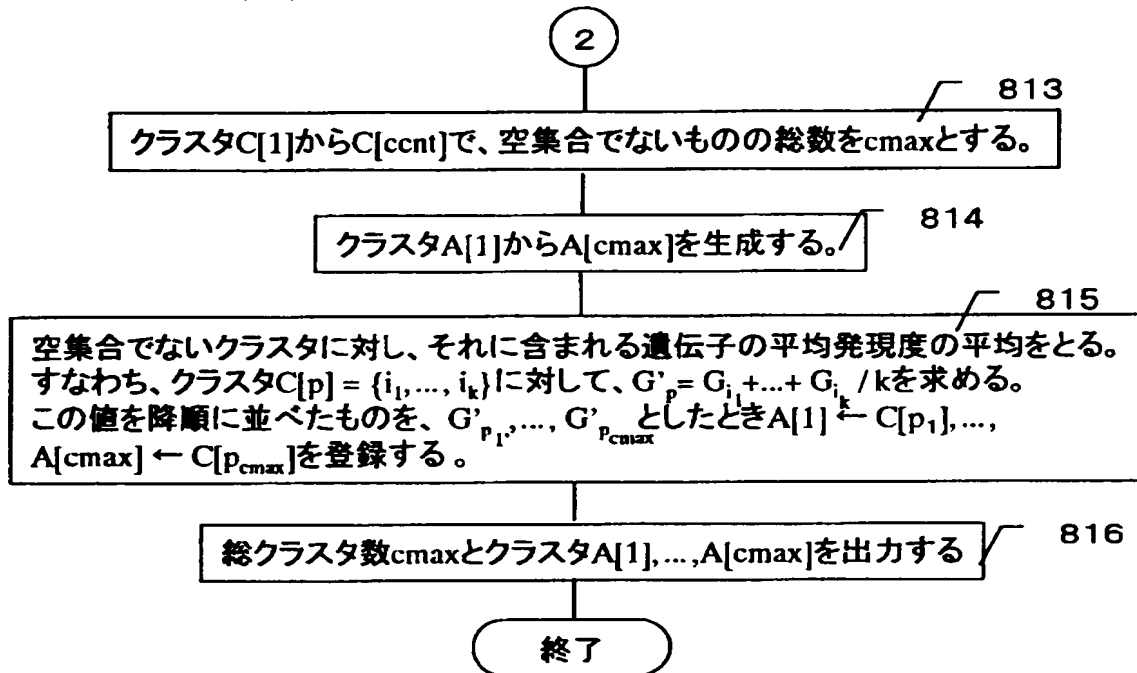
【図 8】

図 8



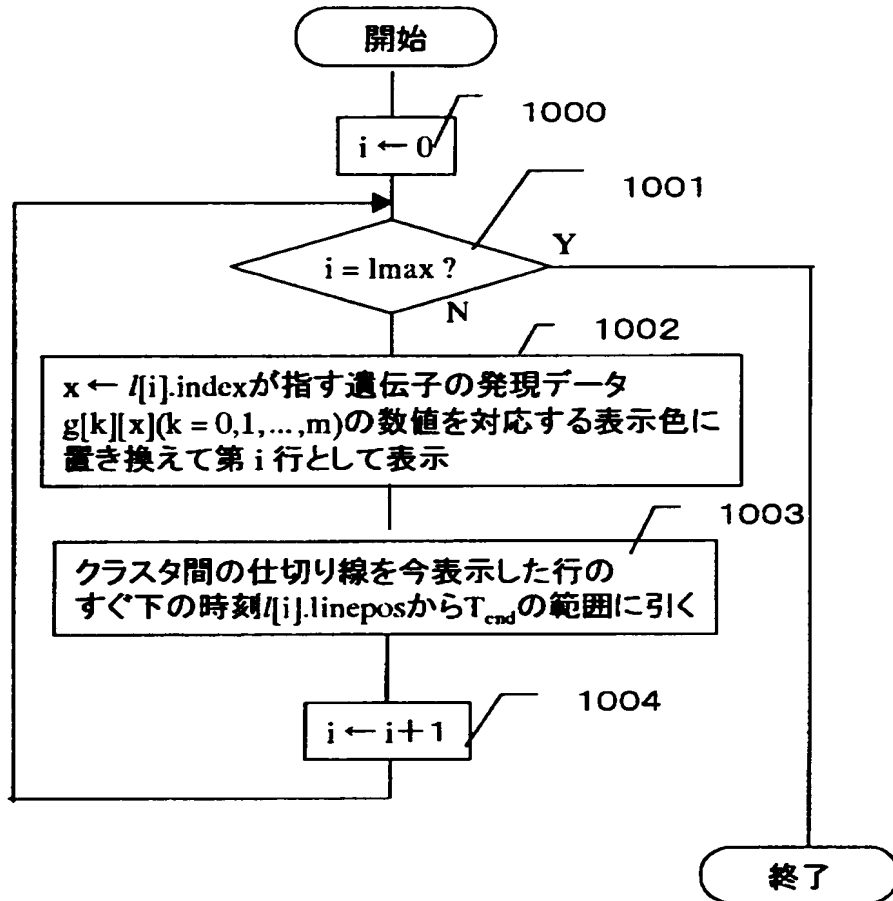
【図 9】

図 9



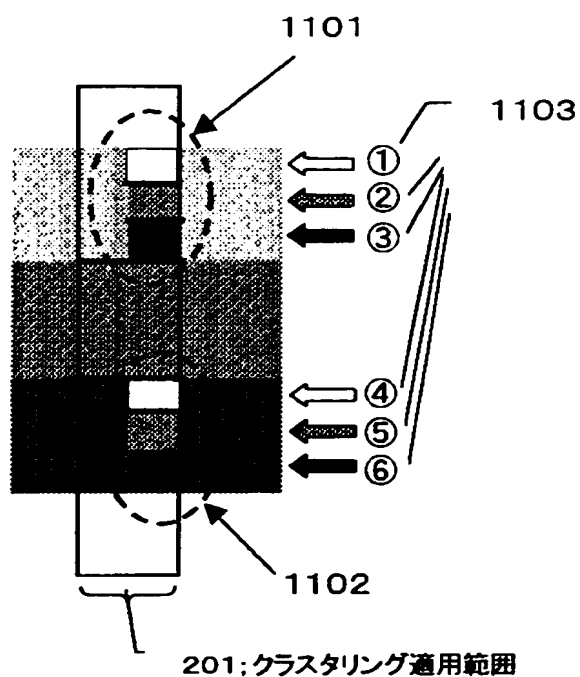
【図 1 0】

図 10



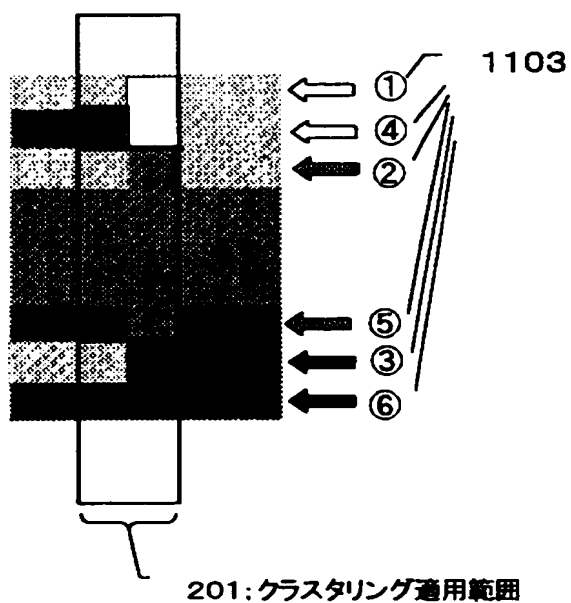
【図 1 1】

図 1 1

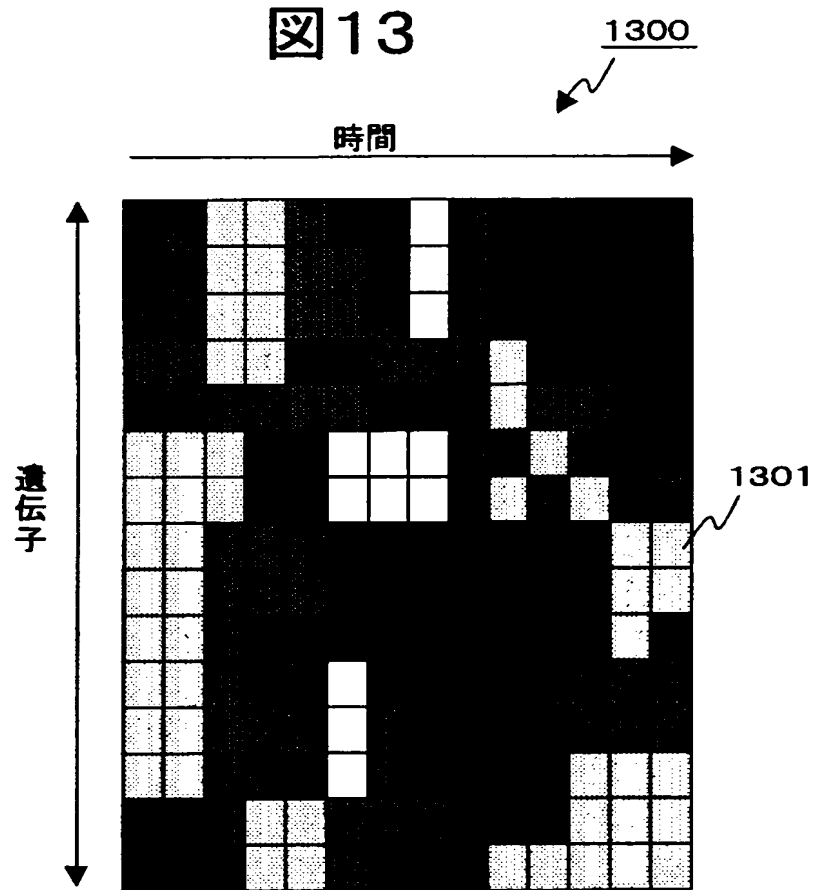


【図 1 2】

図 1 2



【図 1 3】



【書類名】 要約書

【要約】

【課題】 ある時点において異なる遺伝子が同じ機能のために同様に発現しているが、ある時点では別々の役割を持つような場合を見つけ出し、効果的に表示すること。

【解決手段】 複数の遺伝子の時系列発現パターンデータの任意の時間区間を指定するステップと、指定された時間区間における時系列発現パターンデータを予め定めた基準値によってクラスタリングし、さらに同一クラスタ内において基準値を変えながら正または負の時間方向にクラスタリングを繰り返し行い、その結果を予め定めた表示形式で表示するステップとを備えることを特徴とする。

【選択図】 図 4

出 願 人 履 歴 情 報

識別番号 [000233055]

1. 変更年月日	1990年 8月 7日
[変更理由]	新規登録
住 所	神奈川県横浜市中区尾上町6丁目81番地
氏 名	日立ソフトウェアエンジニアリング株式会社